

DARIAH DATA POLICY

Authors:

Françoise Gouzi
Laure Barbot
Matej Ďurčo
Sally Chambers
Toma Tasovac

Contributors:

Erzsébet Tóth-Czifra
Georgios Artopoulos
Panos Constantopoulos
Milena Dobрева
Koraljka Golub
Edward Gray
Tomasz Parkoła
Walter Scholger
Tanja Wissik

Contact: info@dariah.eu

Version 2. November 2024

Table of Content

Table of Content	1
1. Introduction	3
1.1 Target Audiences	4
1.2 DARIAH Resource Types	4
2. DARIAH and Open Science	6
3. Data Producers: how to share your data	7
3.1 Project planning and design phase	8
3.2 Data modelling, standardisation and documentation	10
3.3 Publication	11
4. Data Providers: how to connect your repository to the DARIAH ecosystem	13
4.1 DARIAH Repositories	13
4.2 Trustworthiness and interoperability	14
Conclusion	16
References	17

1. Introduction

We live and work in the age of digital abundance. The massive quantities of available digital content do not, however, automatically translate into easy-to-use research workflows. Digitally enabled, data-informed research in arts and humanities has the potential to transform scientific enquiry and translate its advances into comprehensive solutions, policy actions and socio-economic impact. However, arts and humanities researchers still face a number of challenges finding, accessing, producing and reusing digital resources. For example, managing the volume of data, processing metadata of different quality and granularity, converting metadata for interoperability, clarifying reuse rights for digital objects, choosing the right standards to model data, identifying the best tools to process a myriad of data formats, etc. This policy aims to help address some of these challenges.

The DARIAH Data Policy outlines the basic principles and recommendations for managing and sharing research data within the DARIAH infrastructure and beyond. It can be read alongside/in conjunction with the [DARIAH Service Policy](#) that focuses on service (incl. data services) offer and provision across the network.

Based on the definition provided by the [EU Directive 2019/1024/EU](#)¹, “research data’ means information in a digital form, other than scientific publications, which are collected or produced in the course of scientific research activities and are used as evidence in the research process, or are commonly accepted in the research community as necessary to validate research findings and results”. The term - research data - is highly domain- and context-specific, and has been widely discussed in the Arts and Humanities². Defining and expressing data as ‘data’ is not so common in some disciplinary contexts and oftentimes other concepts such as corpus, content or sources are used. Moreover, as cultural heritage data, originating from cultural heritage institutions, increasingly become humanities research data³, researchers are faced with issues such as insufficient or inconsistent metadata, heterogeneity of formats and legal constraints on accessing and reusing cultural heritage data. Without hindering data sharing, the values of provenance, knowledge management, human and social factors, ethical complexities, temporal dimensions and audience, are some of the many contextual layers to take into account when managing arts and humanities research data.

¹ All links are provided inline but a References Section is also provided at the end of the document

² Edmond, J. , Horsley, N. , Lehmann, J. , & Priddy, M. (2022). *The Trouble With Big Data: How Datafication Displaces Cultural Practices*. London, Bloomsbury Academic. Retrieved December 20, 2023, from <http://dx.doi.org/10.5040/9781350239654>

³ For a DARIAH perspective on cultural heritage data as humanities research data, see Toma Tasovac, Sally Chambers, Erzsébet Tóth-Czifra. *Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper*. 2020. [\(hal-02961317\)](#)

1.1 Target Audiences

The DARIAH Data Policy is primarily aimed at two target audiences: data producers and data providers. In this context, we define these two groups of actors as follows.

A **Data Producer** is an entity (a person, group of persons or an organisation) that creates, generates or collects data for research purposes (see section three).

A **Data Provider** is an entity (a person, group of persons or an organisation) that supplies or makes data available to others.

A data provider may or may not be the original producer of the data and vice versa. Professionals involved in making data available at data-providing institutions include **data stewards, data librarians or research data managers** (see section four).

1.2 DARIAH Resource Types

A large majority of DARIAH resources are produced at DARIAH Partner Institutions in member countries and contributed by their respective national consortia. In addition, there are resources produced in the context of various DARIAH activities, e.g. by [DARIAH Working Groups](#), funded by the [DARIAH Theme calls](#) or by DARIAH and affiliated entities within EU projects.

The term **DARIAH resource** does not indicate direct ownership or control by the DARIAH infrastructure. Instead, it refers to the context of **production and (re)use**: a DARIAH resource is a resource created or (re)used within the DARIAH ecosystem. The same principle of **contextual affiliation** applies to **DARIAH data**: DARIAH does not claim ownership of this data.

When we speak of DARIAH data, we do so in order to foster a common approach to making data, which has been produced and/or provided by DARIAH members, discoverable both within the infrastructure, in the wider context of the [Social Sciences and Humanities \(SSH\) Open Marketplace](#) and the European Open Science Cloud (EOSC). In addition, initiatives such as the [Common European Data Space for Cultural Heritage](#) (led by Europeana) and the [Cultural Heritage Cloud](#), both of which DARIAH is contributing to, enable cultural heritage data (as humanities research data) to be more visible, more easily shared by experts from the GLAM (galleries, libraries, archives, and museums) sector and reused by researchers. In this context, the Collections as Data principles are developed to guide and encourage cultural heritage institutions in publishing their collections, so that they are suitable for computational use.⁴

⁴ See Candela, Gustavo, Chambers, Sally, & Irollo, Alba. (n.d.). A workflow to publish Collections as Data: The case of Cultural Heritage data spaces. Social Sciences & Humanities Open Marketplace. Retrieved 8 November 2024, from <https://marketplace.sshopencloud.eu/workflow/I3lvP6>, and also all the DARIAH

As shown in fig. 1, DARIAH currently recognises 6 main resource types: *publications*, *training materials*, *datasets*, *software*, *services* and *data sources*.

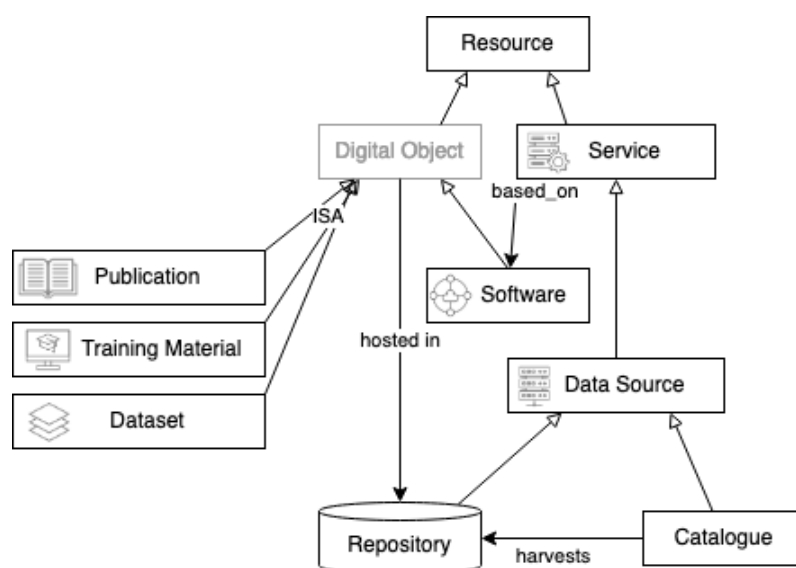


Fig 1. DARIAH resources conceptual view

In the context of the DARIAH Data Policy, we will primarily focus on two specific resource types: *datasets* and *data sources*, which are defined in the table below. Due to their close relationship to data, the definitions of *digital object* and *service* are also added below.

Resource Type	Definition ⁵
Digital Object	Identifiable immaterial items that can be represented as sets of bit sequences, such as datasets, e-texts, images, audio or video items, software, etc., and are documented as single units.
Dataset	A dataset is a digital object or a collection of digital objects that is generally considered a distinctive body of work.
Service	A web application or API which can be accessed online/via the Internet.
Data source	A Data Source is a specific service that exposes metadata about different types of Digital Objects. Examples of data sources are repositories, digital libraries, scientific databases, catalogues, etc.

Section two of this document places DARIAH's data policy in the wider context of DARIAH's involvement in Open Science. Section three addresses data producers, whereas section four addresses data providers.

Campus training materials related to Cultural Heritage:

<https://campus.dariah.eu/tag/cultural-heritage/page/1>

⁵ DARIAH categorises its resources in alignment with the SSH Open Marketplace and EOSC. Definitions are coming from the SSH Open Marketplace [items types](#), the [EOSC Profiles v.5](#) as defined during the EOSC-Future project, the OpenAIRE [types](#) and the [CRMdig v4](#).

2. DARIAH and Open Science

From its foundation, DARIAH has fostered best practices in data management and reuse for arts and humanities researchers, as part of its commitment to Open Science. See, for instance, the DARIAH Impact Case Study: [An Open Science Voice for the Humanities – A Humanities Voice for Open Science](#).

Recent changes in research policy and funding recognise data sharing as a necessary condition for scientific innovation and fostering changes of practice on a systemic level. Open data mandates, and data management plans, are increasingly becoming conditions of research funding both on the European and on national and institutional levels. They impact on the working conditions and practices of all researchers within the broader arts and humanities domain and not only those who identify as digital humanists. Furthermore, datasets are increasingly being described and contextualised through articles in data journals⁶ and as a result being recognised as an innovative form of research output.⁷ This evolution underscores the value of datasets in advancing scholarly discourse. To foster and document innovations in this realm, DARIAH launched in 2024 its own Diamond Open Access overlay journal called [Transformations](#), hosted by Episciences. The journal is dedicated to capturing a wide range of methodological and research activities in arts and humanities, including data collection and processing, annotation and modelling, as well as analysis and interpretation.

DARIAH is a champion of **domain-specific approaches to implementing Open Science**: we aim to help understand and apply generic OS frameworks such as FAIR (Findable, Accessible, Interoperable, Reusable) and CARE (Collective Benefits, Authority to Control, Responsibility, Ethics) to the concrete situations and practices of arts and humanities researchers. The FAIR principles⁸ describe how research outputs should be organised so they can be more easily accessed, understood, exchanged and reused. The CARE principles, on the other hand, describe how data should be treated to ensure that Indigenous governance over the data and its use are respected. “The [CARE principles](#) reflect the crucial role of data in advancing Indigenous innovation and self-determination”. For a DARIAH training resource on how the CARE principles

⁶ See, for instance: Iain Hrynaszkiewicz, Natasha Simons, Azhar Hussain, Rebecca Grant, Simon Goudie. Developing a Research Data Policy Framework for All Journals and Publishers. Data Science Journal, 19 (1). 2020. <https://doi.org/10.5334/dsj-2020-005>; and Jihyun Kim. An Analysis of Data Papers Templates and Guidelines : Types of Contextual Information Described by Data Journals. Science Editing 7 (1) : 16-23. 2020. <https://doi.org/10.6087/kcse.185>

⁷ Toma Tasovac, Laurent Romary, Erzsébet Tóth-Czifra, Rahel C. Ackermann, Daniel Alves, et al.. The Role of Research Infrastructures in the Research Assessment Reform: A DARIAH Position Paper. 2023. <hal-04136772>. <https://hal.science/hal-04136772>

⁸ Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. 2016. <https://doi.org/10.1038/sdata.2016.18>

complement the FAIR principles⁹, see, for instance [Thinking about the CARE Principles in the Digital Humanities](#) on DARIAH-Campus.

A DARIAH Working Group dedicated to [Research Data Management in Arts and Humanities](#) aims to support DARIAH in fulfilling its mission of facilitating access to and reuse of data, as well as supporting the work of the research communities with best practices and methods. It organises workshops to raise awareness, but also produces guidelines translating generic statements on RDM and FAIRness to the specific needs of the A&H communities. The expert members of this group recently released an interactive online publication [Research Data Management for Arts and Humanities: Integrating Voices of the Community](#) which provides an exhaustive overview of the challenges faced by the Arts & Humanities disciplines in relation to RDM.

As part of the SSH Open Cluster (SSHOC) collaboration with other research infrastructures such as CLARIN, CESSDA, E-RIHS or OPERAS, some complementarities may also emerge in terms of the data policies, guidelines and support that cluster members offer their communities. A better overview of these potential overlaps and synergies would enable a further significant and collective contribution to the EOSC from the SSH Open Cluster, to which DARIAH is an active founding member.

Data management practices across Europe remain varied, with some institutions leading the way in terms of commitment and resource allocation, while others are still in the process of organising RDM through training and policy development. In the following section, we will explore how RDM can benefit both researchers and institutions that create, curate or collect data.

3. Data Producers: how to share your data

Effective data management practices are essential for ensuring the accuracy and reliability of research findings. They allow for the preservation of data, enabling others to verify and reproduce the results, which is fundamental to the scientific method. If possible, researchers should be able to replicate analyses using the same datasets to

⁹ Examples of training materials: Elena Giglia, Arnaud Gingold, Iraklis Katsaloulis, Lottie Provost and Francesca Di Donato (2022). FAIR Data in Social Sciences and Humanities. Version 1.0.0. DARIAH-Campus. [Webinar recording]. <https://campus.dariah.eu/id/3fOvyNYHb2Hhg4sQCbtst> and Jennifer Edmond, Matej Ďurčo, Vicky Garnett, Vanessa Hanneschläger, Karlheinz Mörth, Claudia Resch, Basem Saifo, Martina Trognitz, Nienke Van Shaverbeke, Frank Uiterwaal, Dieter Van Uytvanck, Tanja Wissik, Ulrike Wuttke and Jolan Wuyts (2018). Manage, Improve and Open Up Your Research Data. Version 1.0.0. Edited by Vicky Garnett. PARTHENOS Training Suite. [Training module]. <https://training.parthenos-project.eu/sample-page/manage-improve-and-open-up-your-research-and-data/>

validate and build upon existing findings. Systematic data management promotes data sharing and collaboration within the research community. Furthermore, well-managed data ensures its long-term accessibility and usability. This is crucial for preserving research outputs for future generations and avoiding the loss of valuable information.

Preservation of data, reproducibility, long-term accessibility and usability should thus be the main goals to be reached by data producers throughout the research data management process. To achieve this, they may structure and organise their research data in clear steps from a temporal perspective: before, during and after the research project. The data lifecycle model focuses on specific activities for managing data. Each step of the data lifecycle entails concrete tasks of the research process that are interconnected, from one step to the next.¹⁰

The Data Management Plan (DMP) is a central tool of the data management process and must be shared widely and updated regularly. Through this documentation, researchers have to describe and explain clearly the way in which data is handled from the planning, processing to the publishing phase. This document identifies tools, methods and processes ensuring research data are of high-quality, preserved, reproducible and accessible in the long-term, even after the end of the research project.

Online tools such as [DMP Online](#), can help researchers create, review and share Data Management Plans. Furthermore, the ongoing development around machine-actionable DMPs, allowing automated ways to create and update DMPs or connecting them to the data and objects they describe hold, promise for facilitating the management and documentation of data¹¹.

3.1 Project planning and design phase

Creating a data management plan in the early stages of project planning is essential for ensuring the success and long-term impact of the research. It contributes to efficient, legal and ethical data handling, enhances collaboration, and facilitates the reproducibility of research findings. Developing a data management plan helps researchers define the scope and objectives of their project. It forces them to clarify what data will be collected, how it will be managed, and what tools and methodologies will be used. This clarity is essential for maintaining focus and ensuring that the project stays on track.

For instance, when working with human subjects, researchers need to collect appropriate consent forms and must therefore consider how and where their input will

¹⁰ https://clsinfra.io/toolkit_data_sharing/ (Mrugalski, Gouzi, Odebrecht, 2023)

¹¹ See for example the ongoing developments of the OSTRails project led by OpenAire: <https://ostrails.eu/dmps>

be published, as well as how personal data will be processed and protected so as to be compliant with the General Data Protection Regulation (GDPR). In such cases, a DMP can be used to identify and flag various issues and arrive at an agreement with both respondents and project partners. Third-party data providers (such as Cultural Heritage Institutions (CHIs) and/or data repositories) should be involved from the planning phase when the project is still flexible enough to accommodate and sufficiently address preconditions of FAIR and open sharing.

Recommendations

- ❖ Make sure you are compliant with all relevant criteria when estimating the project budget for data management. Consider the costs involved throughout the project lifecycle, including during and after the project, taking into account both the nature of the data and the specific requirements of the project.
- ❖ Use the Heritage Data Reuse Charter template¹² to establish clear and mutually beneficial reuse agreements between CHIs, data providers and researchers during the project planning phase. Use the charter as a guide for aligning expectations and responsibilities regarding data reuse.

Useful resources

- [Planning to meet the costs of managing research data to be FAIR](#) (session 1 of SSHOC / DARIAH RDM WG [workshop](#)).
- [OpenAIRE tool](#): How to identify and assess Research Data Management (RDM) costs?
- [GDPR compliant consent form](#) templates for DH research purposes provided by DARIAH Ethics and Legality in the Digital Arts and Humanities Working Group (ELDAH). This tool enables researchers to correctly approach the processing of personal data in compliance with the General Data Protection Regulation (GDPR).
- [Data management best practices in the Humanities](#) provided by DARIAH RDM Working Group: list of references and resources in RDM.
- [DARIAH Pathfinder to Data Management Best Practices in the Humanities](#): This resource list brings together tools, videos, short articles and other training materials that might be relevant when reflecting on your data management processes both in the immediate context of your research and in their broader disciplinary context.

¹² Erzsébet Tóth-Czifra, Laurent Romary. The Heritage Data Reuse Charter: from principles to research workflows. 2020. ([halshs-02475692](#))

- [Open Data for Humanists, A Pragmatic Guide](#): This resource is aimed at giving practical advice for arts and humanities scholars who are willing to take their first steps in research data management but don't know where to begin.
- A collection of DARIAH-relevant DMPs are available in the Data Management best practices in the Humanities [Zotero bibliography](#).

3.2 Data modelling, standardisation and documentation

When data is shared, it often becomes decoupled from the context of its creation. To ensure that other researchers can understand and effectively reuse the data, this context should be captured in the form of rich and preferably standardised documentation that is accessible to scholars from different backgrounds. This can be a README file or any other supplementary document which includes key information such as provenance, contributors, data formats, curation processes, citation guidelines, limitations and any potential sources of incompleteness (especially in cases where only a subset of the data can be shared due to legal, ethical or other restrictions) or potential bias. Even better, providing this information in standardised formats – using established (meta)data models, vocabularies and ontologies widely recognised in a given field (if any) – will enhance accessibility and usability across disciplines. Information about the technical formats used to store data should also be clearly stated.

Recommendations

- ❖ Use [open file formats](#) to publish your data. Avoid using proprietary file formats; if confronted with proprietary file formats (e.g. as legacy data) convert these to open file formats, whenever possible. However, if the conversion causes information loss, keep the original datasets too (there may be better conversion techniques in the future). Be aware that the storage format can be different from the format that the information is kept in, e.g. you may use a relational database to manage the information, but provide a serialisation in RDF using an established ontology like CIDOC CRM. A detailed list of file formats for archival purposes is available [here](#). In any case, it can be beneficial to store data in more than one file format.¹³
- ❖ Describe the datasets with rich metadata using widely adopted standards such as, for instance, Text Encoding Initiative (TEI), Music Encoding Initiative (MEI), Encoded Archival Description (EAD), Europeana Data Model (EDM), etc. and study metadata interoperability to facilitate as much metadata conversion as possible.

¹³ Archiving textual corpora workflow, <https://marketplace.sshopencloud.eu/workflow/gDzloY>

- ❖ Use existing controlled vocabularies¹⁴ for enriching the data and creating metadata wherever possible.

Useful resources

- [Data model and standards](#) (DARIAH RDM Working Group): best practices and selection of digital preservation-compliant file formats as basis for long-term archiving.
- [CLARIN Standards Information System, especially the](#) recommendations for standards, specifications and the data formats.
- The [LexSeal self-assessment framework](#) translates both the FAIR principles and the 6 core principles of the [Heritage Data Reuse Charter](#) to actual data curation practice and provides an easily understandable self-assessment framework for data quality. Having been co-developed by DARIAH and ELEXIS, it had been designed for lexicographical data sets but curators of other data types can also find relevant know-how in it.
- The [SSH Open Marketplace Workflows](#) have been designed to support researchers in selecting and using the appropriate standards for their particular disciplines and workflows. It will help you to think of data curation as a process that is well-aligned with your specific research lifecycle as well as with your disciplinary contexts.
- [Vocabs services](#) is a DARIAH suite of controlled vocabulary services that allow for collaborative creation, maintenance and publication of vocabularies and taxonomies of any kind.

3.3 Publication

Making data available in a trustworthy data repository is a core requirement of open and FAIR data mandates. By using such repositories, you ensure that humans and machines alike can find your data, and that the data will be preserved over the long term, and therefore remain available for future research or verification.

Trusted repositories typically have established procedures and standards to ensure the integrity of the data they host. They may implement measures such as checksums and data validation checks to prevent corruption or tampering. They also require

¹⁴ An introduction to controlled vocabularies is available in DARIAH-Campus: Ksenia Zaytseva and Matej Ďurčo (2020). Controlled Vocabularies and SKOS. Version 1.1.0. Edited by Matej Ďurčo and Tanja Wissik. DARIAH-Campus. [Training module]. <https://campus.dariah.eu/id/D8d6OrLdpLIGRqBSQDVNO>

documentation and metadata for deposited data. This includes information about how the data was collected, processed, and any relevant contextual details. Proper documentation contributes to better understanding and usability of the data, increasing its quality.¹⁵ Furthermore, reputable data repositories often implement version control mechanisms, allowing users to track changes and updates to datasets over time. This helps maintain a clear historical record of the data, ensuring that users are aware of any modifications or improvements. Finally, trusted repositories will assign persistent identifiers to your datasets, which will facilitate their discoverability and citability.¹⁶

Dynamic data-provision approaches such as Wikidata or a federated set of APIs for programmable corpora (e.g. Dracor¹⁷) are very valuable tools to work with data, but should, nonetheless, be accompanied by regular snapshots to trustworthy repositories (which can be referenced in a stable manner).

Recommendations

- ❖ Clarify all questions regarding intellectual property rights and privacy before making your datasets public.
- ❖ Deposit datasets in a reliable, trustworthy and preferably certified repository (see section 4.2). Always check the archival and long-term preservation policies of the repository (see e.g. Zenodo [policies](#) and [principles](#) and [HAL principles](#))
- ❖ Use an institutional repository, if available and if required by your institutional data sharing policies, or, alternatively, use one of the DARIAH affiliated repositories: [DARIAH HAL collection](#) or [DARIAH Zenodo Community](#).
- ❖ Publish datasets under [CC-BY 4.0 licence](#) by default. Only choose a more restrictive licence if necessary.

Useful resources

- DARIAH affiliated repositories can be found browsing the [DARIAH service catalogue](#).
- [DARIAH's Open Access Guidelines](#): recommendations to improve Open Access to publications in the arts and humanities.

¹⁵ See for example Datasheets for Cultural Heritage: Alkemade, H., Claeysens, S., Colavizza, G., Freire, N., Lehmann, J., Neudecker, C., Osti, G. and van Strien, D. (2023) 'Datasheets for Digital Cultural Heritage Datasets', Journal of Open Humanities Data, 9(1), p. 17. <https://doi.org/10.5334/johd.124>.

¹⁶ See for example CESSDA Persistent Identifier Policy <https://doi.org/10.18448/16.0040> and EOSC PID Policy <https://eosc.eu/advisory-groups/pid-policy-implementation>

¹⁷ Fischer, Frank, et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In Proceedings of DH2019: "Complexities", Utrecht University, <https://doi.org/10.5281/zenodo.4284002>

- [How to find a trustworthy repository for your data?](#) (OpenAIRE): this resource gives support and information on the specificities and criteria of certified repositories for researchers, information managers and other stakeholders.
- [Registry of Research Data Repositories](#) (re3data) covers research data repositories from different academic disciplines.
- [FAIRsharing gives access to a curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies](#) (see section "Certifications and community badges" to find if the repository is certified).
- [Open Data Commons](#) proposes a set of legal tools and open licences ([PDDL](#) or [ODC-By](#)) to help you publish, provide and use open data and [Open Software Licence](#) for Softwares.
- [This paper](#) developed within the framework of the CENDARI project outlines key steps of sustaining outputs of research projects.

In this section addressed to Data Producers, we observed how data management is crucial to ensure that research data is: archived in a sustainable repository, easy to find, understandable, and (re)usable. The different best practices and resources should help researchers and institutions to consolidate the research lifecycle.

4. Data Providers: how to connect your repository to the DARIAH ecosystem

While the previous section focused on DARIAH recommendations to data producers (e.g. researchers and research institutions), this section is aimed at data providers. It offers guidance to individuals and institutions responsible for data services, outlining recommendations for those looking to enhance their connections with the DARIAH ecosystem.

4.1 DARIAH Repositories

As a result of the complex institutional landscape of European Research Infrastructures, not all the datasets registered or hosted in the DARIAH repositories are automatically considered DARIAH data. For example, an institutional and discipline-agnostic

repository, such as the French [Open Archive HAL](#), is onboarded as a DARIAH service. However, not all of the scientific papers or other objects in this repository (i.e in Physics or Life sciences) should automatically be considered as DARIAH data.

Repositories are considered as well as other data sources, such as scientific databases, journal and publisher archives, research information systems (CRIS) and aggregators (e.g. [OpenAire Research Graph](#) or [re3data](#)). Most of the DARIAH data sources are repositories declared by DARIAH National Coordinators as part of their national reporting. In the context of this data policy, we focus exclusively on repositories.

Recommendations

- ❖ Register the repository in the Marketplace and tag it properly (see Guidelines for [Contributing to the SSH Open Marketplace](#)).
- ❖ Register your repository with OpenAIRE by making its metadata compliant with the [OpenAire Guidelines](#).

Metadata that is compliant with the OpenAIRE Guidelines will be included in the [OpenAire Research Graph](#), and also in the [EOSC EU Node Resource Hub](#).

4.2 Trustworthiness and interoperability

Over the last 10 years, many universities and research centres have developed data repositories allowing long-term access to datasets in a trustworthy environment. However, it is not always easy to evaluate the quality of these repositories¹⁸. To this end, some certification instruments for digital repositories have been developed, a prominent one being the [CoreTrustSeal \(CTS\)](#). At the same time, there are also data repositories that have no certification but have earned trustworthiness through many years of reliable operation by trustworthy providers and large user bases, like [Zenodo](#).

Aggregation systems frequently prioritise simplified metadata schemas to ensure interoperability, which can lead to a reductionist representation of the data's rich context and detail. The "risk of losing the thick description"¹⁹ of humanities research data is therefore a key challenge for DARIAH data providers to keep in mind. Data providers should prioritise the use of robust metadata schemas that capture detailed information on provenance, ownership and conditions of reuse. At the same time, they should

¹⁸ Disappearing repositories -- taking an infrastructure perspective on the long-term availability of research data <https://doi.org/10.48550/arXiv.2310.06712>

¹⁹ Tóth-Czifra, E. (2020). The Risk of Losing the Thick Description : Data Management Challenges Faced by the Arts and Humanities in the Evolving FAIRData Ecosystem. In Edmond, J. (Ed.), Digital Technology and the Practices of Humanities Research. Open Book Publishers.
<https://www.openbookpublishers.com/books/10.11647/obp.0192/chapters/10.11647/obp.0192.10>

ensure that the user interface of their data services remains user-friendly, allowing data producers to easily register and manage their assets. Balancing rich metadata with simplicity in design is essential for effective data management.

Recommendations

- ❖ Study and follow the [Core Trust Seal](#) Repositories requirements. They describe the characteristics required to be a trustworthy repository for digital data and metadata²⁰.
- ❖ Implement the [OAI-PMH protocol](#) as a default mechanism to facilitate harvesting of the metadata by third party aggregators, in order to promote dissemination and thus findability of the resources.
- ❖ Expose metadata adhering to the [DataCite Metadata Schema](#) adopted by OpenAIRE as a minimal schema.
- ❖ Expose metadata in multiple formats according to the requirements of relevant aggregators, to optimally convey existing information.
- ❖ Use specialised schemas²¹ tailored to domain-specific needs internally in order to capture as much as possible relevant and available information (e.g. regarding the context of creation, provenance, conditions of reuse).
- ❖ Use [persistent identifiers \(PIDs\)](#) to identify datasets in the repository following one of the established PID systems: e.g. DOI, Handle, ARK, URN.
- ❖ Honour [Data Citation Principles](#) (Importance, Credit and Attribution, Evidence, Unique Identification, Access, Persistence, Specificity and Verifiability, Interoperability and Flexibility) enabling data reuse.
- ❖ If providing resources with restricted access, adopt Federated Identity Management using technologies such as [Security Assertion Markup Language](#) (SAML) and [OpenID Connect](#) (OIDC) and integrate with relevant Authentication and Authorisation Infrastructures (AAI) most notably eduGAIN and/or DARIAH-AAI (as part of EOSC-AAI Federation).

²⁰ In the DARIAH context, [GAMS](#) or [TextGrid](#) are examples of repositories carrying the CoreTrustSeal. See also the SSHOC support action to improve repositories trust and quality insurance: <https://sshopencloud.eu/sshoc-trusted-repositories>

²¹ See for example how the Swedish National Data Service developed “subject-specific profiles (...) developed according to domain-specific metadata standards and international research infrastructures”: <https://snd.se/en/describe-and-share-data/metadata>

Useful resources

- [Introduction to Persistent Identifiers](#): DARIAH-CAMPUS webinar focusing on 'Persistent Identifiers' (PIDs) and basic concepts of referencing objects.
- [CoreTrustSeal Requirements 2023-2025](#): describes the characteristics required to be a trustworthy repository for digital data and metadata.
- [How to improve the quality of your repository - SSHOC and certification of repositories](#) (15 min. Webinar): this resource is aimed at staff of data repositories interested in certification.
- [ePIC](#) - Persistent Identifiers for eResearch: consortium of European partners in order to provide PID services for the European Research Community, based on the Handle system.
- [OpenDOAR](#): global Directory of Open Access Repositories, hosting repositories that provide free, open access to academic outputs and resources. See the Inclusion Criteria if you want your repository to be listed in OpenDOAR.
- [Re3data](#): gives the requirements to be registered as a research data repository.

Conclusion

The DARIAH Data Policy aims at supporting data producers and data providers by providing the basic principles and recommendations for managing and sharing research data within the DARIAH infrastructure and beyond.

Grounded in the [DARIAH Open Science context](#) and guided by the main DARIAH resource types, the DARIAH Data Policy advocates for efficient data management practices promoting data sharing and collaboration within the research community and among the whole research life cycle, from project planning and standardisation to publication. Alongside the DARIAH Service Policy, and the upcoming Guidelines for handling DARIAH resources, the DARIAH Data Policy contributes to the DARIAH resources management framework of the infrastructure.

The DARIAH Data Policy guides researchers and institutions along some of the recommended paths to be better connected to the DARIAH ecosystem, and as thus, more visible in the European Digital Arts and Humanities landscape.

This second version of the DARIAH Data Policy is the result of a collective effort from the DARIAH community. However, it is still considered a living document that will need regular updates.

References

ACDH-CH/DARIAH. 2024. Vocab services. Skosmos version 2.16. Accessed in November 2024: <https://vocab.dariah.eu/en/>

Alkemade, H., Claeysens, S., Colavizza, G., Freire, N., Lehmann, J., Neudecker, C., Osti, G. and van Strien, D. 2023. Datasheets for Digital Cultural Heritage Datasets, Journal of Open Humanities Data, 9(1), p. 17. Accessed in November 2024: <https://doi.org/10.5334/johd.124>.

Australian Research Data Commons. 2022. CARE Principles. Accessed in November 2024: <https://ardc.edu.au/resource/the-care-principles/>

Buelinckx, E., Gautschi, R., Gelati, F., & Király, P. 2023. Section 3.6.2. Data archiving and (digital) long-term archiving. Research Data Management for Arts and Humanities. GitLab Repository. Accessed in November 2024: <https://gitlab-ce.rrz.uni-hamburg.de/uahh-digitale-dienste/rdm-for-arts-and-humanities/-/blob/master/part-03.md#362-data-archiving-and-digital-long-term-archiving>

Candela, Gustavo, Chambers, Sally, & Irollo, Alba. (n.d.). A workflow to publish Collections as Data: The case of Cultural Heritage data spaces. Social Sciences & Humanities Open Marketplace. Accessed in November 2024: <https://marketplace.sshopencloud.eu/workflow/I3jvP6>

CCSD. 2024. HAL Science Ouverte. Accessed in November 2024: <https://hal.science/>

CCSD. 2024. HAL Principles. Accessed in November 2024: <https://about.hal.science/en/principles/>

CERN. Zenodo. 2024. Accessed in November 2024: <https://zenodo.org/>

CERN. Zenodo General Policies v1.0. Accessed in November 2024: <https://about.zenodo.org/policies/>

CESSDA. 2017. CESSDA ERIC Persistent Identifier Policy. Version 1.0 Accessed in November 2024: <https://doi.org/10.18448/16.0040>

CLARIN. 2024. CLARIN Standards Information System. Accessed in November 2024: <https://clarin.ids-mannheim.de/standards/>

CoreTrustSeal. 2024. CoreTrustSeal. Accessed in November 2024: <https://www.coretrustseal.org/>

CoreTrustSeal Standards and Certification Board. 2022. CoreTrustSeal Trustworthy Digital Repositories Requirements 2023-2025 Extended Guidance (V01.00). Zenodo. Accessed in November 2024: <https://doi.org/10.5281/zenodo.7051096>

Creative Commons. 2024. Accessed in November 2024: <https://creativecommons.org/licenses/by/4.0/>

CRM-CIDOC. 2022. Definition of CRMdig v4.0. Accessed in November 2024: <https://cidoc-crm.org/crmvig/ModelVersion/version-4.0>

DARIAH. 2019. Cultural Heritage Data Reuse Charter: the Mission Statement. Accessed in November 2024: <https://datacharter.hypotheses.org/mission-statement>

DARIAH. 2024. DARIAH HAL Collection. Accessed in November 2024: <https://shs.hal.science/DARIAH/>

DARIAH. 2024. DARIAH Zenodo Community. Accessed in November 2024: <https://zenodo.org/communities/dariah/records?q=&l=list&p=1&s=10&sort=newest>

DARIAH. 2024. Tools and Services Catalogue. Accessed in November 2024: <https://www.dariah.eu/tools-services/tools-and-services/>

DARIAH-EU. 2023. Research Data Management (Working Group). Accessed in November 2024: <https://www.dariah.eu/activities/working-groups/research-data-management/>

DARIAH RDM Working Group. 2024. Data Management best practices in the Humanities. DMPs Collection. Zotero Group Library. Accessed in November 2024: https://www.zotero.org/groups/2427138/data_management_best_practices_in_the_humanities/collections/3Q44QWAH

DARIAH RDM Working Group. 2024. Data management best practices in the Humanities. Zotero Group Library. Accessed in November 2024: https://www.zotero.org/groups/2427138/data_management_best_practices_in_the_humanities

Davidson, J., Helbig, K., Montesanti, A., O'Connor, R., Schuster, K., Watteler, O., Weller, K., & Whyte, A. 2024. Train-the-Trainer Research Data Management Bootcamp. Accessed in November 2024: <https://campus.dariah.eu/resource/posts/train-the-trainer-research-data-management-bootcamp>

Digital Curation Centre. 2024. DMP Online. Accessed in November 2024: <https://dmponline.dcc.ac.uk/>

Dutch Digital Heritage Network. 2024. Persistent Identifier Guide. Accessed in November 2024: <https://www.pidwijzer.nl/en>

Edmond, J. & Morselli, F. 2020. Sustainability of Digital Humanities Projects as a Publication and Documentation Challenge. Trinity's Access to Research Archive. Accessed in November 2024: <https://www.tara.tcd.ie/handle/2262/91506>

Edmond, J., & Tóth-Czifra, E. 2018. Open Data for Humanists, A Pragmatic Guide. Zenodo. Accessed in November 2024: <https://doi.org/10.5281/zenodo.2657248>

ePIC. 2024. ePIC - Persistent Identifiers for eResearch. Accessed in November 2024: <http://www.pidconsortium.net/>

ELDAH (Ethics and Legality in Digital Arts and Humanities). 2021. Welcome to the DARIAH ELDAH Consent Form Wizard (CFW)! Accessed in November 2024: <https://consent.dariah.eu/>

EOSC. 2024. PID Policy and Implementation. Accessed in November 2024: <https://eosc.eu/advisory-groups/pid-policy-implementation>

EUR-Lex. 2019. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast). Accessed in November 2024: <https://eur-lex.europa.eu/eli/dir/2019/1024/oj>

Europeana. 2024. Common European Data Space for Cultural Heritage. Accessed in November 2024: <https://www.dataspace-culturalheritage.eu/en>

European Commission. European Open Science Cloud EU Node website. Accessed in November 2024: <https://open-science-cloud.ec.europa.eu/>

Ghent University. 2024. RDM explained, What is Research Data Management (RDM)? <https://www.ugent.be/en/research/openscience/datamanagement/why/rdm-explained.htm> and Preparing a Data Management Plan (DMP). Accessed in November 2024: <https://www.ugent.be/en/research/openscience/datamanagement/before-research/datamanagementplan.htm>

JISC. 2024. OpenDOAR. Accessed in November 2024: <https://v2.sherpa.ac.uk/opensdoar/about.html>

Kalman, T. 2022. Introduction to Persistent Identifiers. DARIAH-Campus. [Webinar recording]. Accessed in November 2024: <https://campus.dariah.eu/resource/posts/introduction-to-persistent-identifiers>

Ksenia Zaytseva and Matej Ďurčo. 2020. Controlled Vocabularies and SKOS. Version 1.1.0. Edited by Matej Ďurčo and Tanja Wissik. DARIAH-Campus. [Training module]. Accessed in November 2024: <https://campus.dariah.eu/id/D8d6OrLdpLIGRqBSQDVN0>

List of open file formats. 2024. Wikipedia. Accessed in November 2024: https://en.wikipedia.org/wiki/List_of_open_file_formats

Martone, M. (ed.). 2014. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. San Diego CA: FORCE11. Accessed in November 2024: <https://doi.org/10.25490/a97f-egykh>

O'Donnell, D. 2021. Thinking about the CARE Principles in the Digital Humanities. DARIAH-CAMPUS. Accessed in November 2024: <https://campus.dariah.eu/resource/posts/thinking-about-the-care-principles-in-the-digital-humanities>

OASIS standards consortium. 2020. SAML Wiki. Accessed in November 2024: <https://wiki.oasis-open.org/security/FrontPage>

OpenAIRE. 2020. Guides for Researchers. How to identify and assess Research Data Management (RDM) costs. Accessed in November 2024: <https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs>

OpenAIRE. 2020. How to find a trustworthy repository for your data? Accessed in November 2024: <https://www.openaire.eu/find-trustworthy-data-repository>

OpenAIRE. 2024. OpenAIRE Graph. Accessed in November 2024: <https://graph.openaire.eu/>

OpenAIRE. 2022. OpenAIRE Guidelines. Accessed in November 2024: <https://guidelines.openaire.eu/en/latest/>

OpenAIRE. 2022. OpenAIRE Guidelines. Use of DataCite. Accessed in November 2024: https://guidelines.openaire.eu/en/latest/data/use_of_datacite.html#use-of-datacite

OpenID. 2024. What are OpenID Specifications. Accessed in November 2024: <https://openid.net/developers/specs/>

Open Knowledge. 2024. Open Data Commons Legal tools for Open Data. Accessed in November 2024: <https://opendatacommons.org/licenses/pddl/>

re3data.org. 2024. Registry of Research Data Repositories (Search Page). Accessed in November 2024: <https://doi.org/10.17616/R3D>

re3data.org. 2024. Suggest a repository. Accessed in November 2024: <https://www.re3data.org/suggest>

Research Data Management for Arts and Humanities. 2022. Research Data Management for Arts and Humanities: Integrating Voices of the Community. Accessed in November 2024:

<https://gitlab-ce.rrz.uni-hamburg.de/uahh-digitale-dienste/rdm-for-arts-and-humanities>

Romary, L., Tóth-Czifra, E. 2019. Open Access guidelines for the arts and humanities: Recommendations by the DARIAH European research infrastructure consortium. Accessed in November 2024: <https://shs.hal.science/halshs-02106332>

Rosen, L. 2023. Open Software License. Wikipedia. Accessed in November 2024: https://en.wikipedia.org/wiki/Open_Software_License

SSHOC. 2020. SSHOC Webinar: How to improve the quality of your repository - SSHOC and certification of repositories. Zenodo. Accessed in November 2024: <https://zenodo.org/records/3774396>

Social Sciences and Humanities Open Cloud project (SSHOC). 2023. Help grow the SSH Open Marketplace? Accessed in November 2024: <https://marketplace.sshopencloud.eu/contribute/overview>

Social Sciences and Humanities Open Cloud project (SSHOC). 2024. SSH Open MarketPlace. Accessed in November 2024: <https://marketplace.sshopencloud.eu>

Social Sciences and Humanities Open Cloud project (SSHOC). 2024. [SSH Open Marketplace Workflows](https://marketplace.sshopencloud.eu/search?categories=workflow). Accessed in November 2024: <https://marketplace.sshopencloud.eu/search?categories=workflow>

Social Sciences and Humanities Open Cloud project (SSHOC). SSHOC Trusted Repositories. Accessed in November 2024: <https://sshopencloud.eu/sshoc-trusted-repositories>

Tasovac, T., Chambers, S., Tóth-Czifra, E. 2020. Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper. Accessed in November 2024: <https://hal.science/hal-02961317>

Tasovac, T., Romary, L., Tóth-Czifra, E. & Marinski, I. 2021. Lexicographic Data Seal of Compliance. DARIAH. Accessed in November 2024: <https://hal.science/hal-03344267>

The FAIRsharing team Data Readiness Group University of Oxford. 2024. FAIRsharing.org - standards, databases, policies. Accessed in November 2024: <https://fairsharing.org/>

Toolkit for Data Sharing (CLSIInfra. Mrugalski, Gouzi, Odebrecht). 2023. Accessed in November 2024: https://clsinfra.io/toolkit_data_sharing/

Tóth-Czifra, E. 2021. DARIAH Impact Case Study series: An Open Science voice for the Humanities - a Humanities voice for Open Science. Accessed in November 2024: <https://doi.org/10.5281/zenodo.5863209>

Tóth-Czifra, E. 2024. DARIAH Pathfinder to Data Management Best Practices in the Humanities. DARIAH-CAMPUS. Accessed in November 2024: <https://campus.dariah.eu/resource/posts/dariah-pathfinder-to-data-management-best-practices-in-the-humanities>

Whyte, A., Davidson, J., O'Connor, R., & Montesanti, A. 2021. Planning to meet the costs of managing research data to be FAIR. DARIAH-CAMPUS. Accessed in November 2024: <https://doi.org/10.5281/zenodo.4518901>